



# EP3260: Machine Learning Over Networks

## Lecture 3: Centralized Convex ML

(part 2)

Hossein S. Ghadikolaei

Division of Network and Systems Engineering  
School of Electrical Engineering and Computer Science  
KTH Royal Institute of Technology, Stockholm, Sweden  
<https://sites.google.com/view/mlons2023/home>

January 2023

# Learning outcomes

- Recap of (deterministic) iterative algorithms for convex optimization
- Stochastic optimization
- Variance reduction techniques
- Convergence analysis

# Outline

1. Basic definitions and properties
2. Problem Statement
3. Fundamental Lemmas and Assumptions
4. Convergence Results for SG
5. Variance Reduction Techniques
6. Supplements

# Recap of Lecture 2 and beyond!

## Smooth problems ( $L$ -smooth, $\mu$ -strong convexity)

Gradient descent: minimize $_{\mathbf{w} \in \mathbb{R}^d}$   $f(\mathbf{w})$ ,  $\mathcal{O}(1/k)$  for convex

Projected gradient descent: minimize $_{\mathbf{w} \in \mathcal{W}}$   $f(\mathbf{w})$ ,  $\mathcal{O}(1/k)$  for convex

Steepest descent: minimize $_{\mathbf{w} \in \mathcal{W}}$   $f(\mathbf{w})$ , large  $L/\mu$ ,  $\mathcal{O}(1/k)$  for convex

Newton's methods: minimize $_{\mathbf{w} \in \mathcal{W}}$   $f(\mathbf{w})$ , large  $L/\mu$

Acceleration methods: minimize $_{\mathbf{w} \in \mathcal{W}}$   $f(\mathbf{w})$ , large  $L/\mu$ ,  $\mathcal{O}(1/k^2)$  for convex

## Nonsmooth problems

Subgradient methods: minimize $_{\mathbf{w} \in \mathbb{R}^d}$   $f(\mathbf{w})$ ,  $\mathcal{O}(1/k)$  for convex

Proximal methods: minimize $_{\mathbf{w} \in \mathbb{R}^d}$   $g(\mathbf{w}) + h(\mathbf{w})$ ,  $\mathcal{O}(1/k)$  for smooth  $f$

Accelerated proximal methods: minimize $_{\mathbf{w} \in \mathbb{R}^d}$   $g(\mathbf{w}) + h(\mathbf{w})$ , convex  $h$ ,  $\kappa = L/\mu$

$$\text{update: } \mathbf{w}_{k+1} = \text{prox}_{\alpha_k h}(\mathbf{v}_k - \alpha_k \nabla g(\mathbf{v}_k))$$

$$\text{momentum from prev. iteration: } \mathbf{v}_{k+1} = \mathbf{w}_{k+1} + \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}(\mathbf{w}_{k+1} - \mathbf{w}_k)$$

# Outline

1. Basic definitions and properties
2. Problem Statement
3. Fundamental Lemmas and Assumptions
4. Convergence Results for SG
5. Variance Reduction Techniques
6. Supplements

# Basic definitions

Convexity for differentiable function:

$$\nabla f(\mathbf{w}_1)^T (\mathbf{w}_2 - \mathbf{w}_1) \leq f(\mathbf{w}_2) - f(\mathbf{w}_1)$$

Strongly convexity:

$$f(\mathbf{w}_2) \geq f(\mathbf{w}_1) + \nabla f(\mathbf{w}_1)^T (\mathbf{w}_2 - \mathbf{w}_1) + \frac{\mu}{2} \|\mathbf{w}_2 - \mathbf{w}_1\|_2^2$$

Smoothness:

$$f(\mathbf{w}_2) \leq f(\mathbf{w}_1) + \nabla f(\mathbf{w}_1)^T (\mathbf{w}_2 - \mathbf{w}_1) + \frac{L}{2} \|\mathbf{w}_2 - \mathbf{w}_1\|_2^2$$

Bounded error for initial guess:  $\mathbb{E} [\|\mathbf{w}_1 - \mathbf{w}^*\|_2] \leq R$

Lipschitz continuity (bounded gradients)

$$\begin{aligned} \|\mathbf{w}\|_2 \leq D &\Rightarrow \|\nabla f(\mathbf{w})\|_2 \leq B \\ \text{or } \|\mathbf{w}_1\|_2, \|\mathbf{w}_2\|_2 \leq D &\Rightarrow |f(\mathbf{w}_2) - f(\mathbf{w}_1)| \leq B \|\mathbf{w}_2 - \mathbf{w}_1\|_2 \end{aligned}$$

# Example

Consider Human Activity Recognition Using Smartphones dataset

$$\{(\mathbf{x}_i, y_i)\}_{i \in [N]}$$

inputs: accelerometer and gyroscope sensors

output: moving (e.g., walking, running, dancing) or not (sitting or standing)

Consider logistic ridge regression: minimize  $f(\mathbf{w}) = \frac{1}{N} \sum_{i \in [N]} f_i(\mathbf{w}) + \lambda \|\mathbf{w}\|_2^2$

where  $f_i(\mathbf{w}) = \log(1 + \exp\{-y_i \mathbf{w}^T \mathbf{x}_i\})$

For classification, we can use the solution  $\mathbf{w}^*$  and compute  $\text{sign}(\mathbf{w}^{*T} \mathbf{x})$

## HW 2.1:

- 1) Is  $f$  Lipschitz continuous? If so, find a small  $B$ ?
- 2) Is  $f_i$  smooth? If so, find a small  $L$  for  $f_i$ ? What about  $f$ ?
- 3) Is  $f$  strongly convex? If so, find a high  $\mu$ ?

# Outline

1. Basic definitions and properties
- 2. Problem Statement**
3. Fundamental Lemmas and Assumptions
4. Convergence Results for SG
5. Variance Reduction Techniques
6. Supplements



# Setting

- **Batch GD:** Let  $f(\mathbf{w}) = \frac{1}{N} \sum_{i \in [N]} f_i(\mathbf{w})$

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \alpha_k \nabla f(\mathbf{w}_k) = \mathbf{w}_k - \frac{\alpha_k}{N} \sum_{i \in [N]} \nabla f_i(\mathbf{w}_k)$$

- **Stochastic gradient (SG) methods:**

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \alpha_k g(\mathbf{w}_k; \zeta_k) = \mathbf{w}_k - \alpha_k \hat{\nabla} f(\mathbf{w}_k)$$

$\zeta_k \in [N]$ , and  $g(\mathbf{w}_k; \zeta_k)$  is a noisy version (“estimation”) of  $\nabla f(\mathbf{w}_k)$ .

Method	Per iteration cost	# iterations
GD	Expensive (usually linear in $N$ )	Usually few
SG	Very cheap, independent of $N$	Many

**Main tradeoff:** Per-iteration cost vs per-iteration improvement

# Motivations for SG

Good theoretical guarantees: Consider strongly convex smooth  $f$ , then

- GD:  $f(\mathbf{w}_k) - f(\mathbf{w}^*) \leq \mathcal{O}(\rho^k)$ ,  $\rho \in (0, 1)$ , so  $N \log(1/\epsilon)$  total work for  $\epsilon$ -optimality
- SG (basic version):  $\mathbb{E}[f(\mathbf{w}_k) - f(\mathbf{w}^*)] \leq \mathcal{O}(1/k)$ , so  $1/\epsilon$  total work for  $\epsilon$ -optimality
- Compare  $N \log(1/\epsilon)$  to  $1/\epsilon$  for large  $N$

Heavy computation

- Large scale optimization,  $N \rightarrow \infty$ , large matrix inversion

Heavy communication

- Bandwidth-limited distributed optimization

Privacy

- Revealing only a noisy gradient information

Nonconvex optimization and saddle points

# Generic SG algorithm for decentralized optimization

## A generic SG algorithm

```
Initialize  $\mathbf{w}_1$ 
for  $k = 1, 2, \dots$ , do
    Generate a realization of the random variable  $\zeta_k$ 
    Compute a stochastic vector  $g(\mathbf{w}_k; \zeta_k)$ 
    Choose step-size  $\alpha_k > 0$ 
    Update  $\mathbf{w}_{k+1} \leftarrow \mathbf{w}_k - \alpha_k g(\mathbf{w}_k; \zeta_k)$ 
end for
```

- Problem: minimize  $f(\mathbf{w}) = \frac{1}{N} \sum_{i \in [N]} f_i(\mathbf{w})$ .

- Examples of stochastic vector

Gradient for one sample:  $\nabla f_{\zeta_k}(\mathbf{w}_k)$

Gradient for a mini-batch:  $\frac{1}{N_k} \sum_{i \in [N_k]} \nabla f_{\zeta_k, i}(\mathbf{w}_k)$

Preconditioned mini-batch gradient:  $H_k \frac{1}{N_k} \sum_{i \in [N_k]} \nabla f_{\zeta_k, i}(\mathbf{w}_k)$

# Outline

1. Basic definitions and properties
2. Problem Statement
3. Fundamental Lemmas and Assumptions
4. Convergence Results for SG
5. Variance Reduction Techniques
6. Supplements

# Smoothness

Observe that  $\mathbf{w}_{k+1}$  depends only on  $\zeta_k$ , and assume i.i.d.  $(\zeta_k)_k$

$\mathbb{E}_{\zeta_k}[f(\mathbf{w}_{k+1})]$ : expectation of  $f(\mathbf{w}_{k+1})$  wrt the distribution of  $\zeta_k$  only

$f$  being  $L$ -smooth implies that the generic SG algorithm satisfies for all  $k \in \mathbb{N}$

$$\begin{aligned} \mathbb{E}_{\zeta_k}[f(\mathbf{w}_{k+1})] - f(\mathbf{w}_k) \leq \\ \underbrace{-\alpha_k \nabla f(\mathbf{w}_k)^T \mathbb{E}_{\zeta_k}[g(\mathbf{w}_k; \zeta_k)]}_{\text{expected decrease}} + \underbrace{\frac{1}{2} \alpha_k^2 L \mathbb{E}_{\zeta_k}[\|g(\mathbf{w}_k; \zeta_k)\|_2^2]}_{\text{noise}} \end{aligned}$$

If  $g(\mathbf{w}_k; \zeta_k)$  is an unbiased estimate of  $\nabla f(\mathbf{w}_k)$ , then

$$\mathbb{E}_{\zeta_k}[f(\mathbf{w}_{k+1})] - f(\mathbf{w}_k) \leq -\alpha_k \|\nabla f(\mathbf{w}_k)\|_2^2 + \frac{1}{2} \alpha_k^2 L \mathbb{E}_{\zeta_k}[\|g(\mathbf{w}_k; \zeta_k)\|_2^2] \quad (1)$$

## Some useful assumptions

- The sequence  $\{\mathbf{w}_k\}$  is contained in an open set over which  $f$  is bounded below by a scalar  $f_{\inf}$
- There exist scalars  $c_0 \geq c > 0$  s.t. for all  $k \in \mathbb{N}$

$$\nabla f(\mathbf{w}_k)^T \mathbb{E}_{\zeta_k} [g(\mathbf{w}_k; \zeta_k)] \geq c \|\nabla f(\mathbf{w}_k)\|_2^2 \quad (2a)$$

$$\|\mathbb{E}_{\zeta_k} [g(\mathbf{w}_k; \zeta_k)]\|_2 \leq c_0 \|\nabla f(\mathbf{w}_k)\|_2 \quad (2b)$$

- There exist scalars  $M \geq 0$  and  $M_V \geq 0$  s.t. for all  $k \in \mathbb{N}$

$$\text{Var}_{\zeta_k} [g(\mathbf{w}_k; \zeta_k)] \leq M + M_V \|\nabla f(\mathbf{w}_k)\|_2^2 \quad (3)$$

For unbiased gradient estimator:  $c = c_0 = 1$

(2) and (3) imply (HW 2.2: find  $M_G$ .)

$$\mathbb{E}_{\zeta_k} [\|g(\mathbf{w}_k; \zeta_k)\|_2^2] \leq M + M_G \|\nabla f(\mathbf{w}_k)\|_2^2$$

## An important tradeoff

Generic SG algorithm on  $L$ -smooth function satisfies

$$\begin{aligned}\mathbb{E}_{\zeta_k} [f(\mathbf{w}_{k+1})] - f(\mathbf{w}_k) &\leq -c\alpha_k \|\nabla f(\mathbf{w}_k)\|_2^2 + \frac{1}{2}\alpha_k^2 L \mathbb{E}_{\zeta_k} [\|g(\mathbf{w}_k; \zeta_k)\|_2^2] \\ &\leq -\left(c - \frac{1}{2}\alpha_k L M_G\right) \alpha_k \|\nabla f(\mathbf{w}_k)\|_2^2 + \frac{1}{2}\alpha_k^2 L M \quad (4)\end{aligned}$$

*Proof:* see the board

Convergence of SG depends on the balance between blue and red terms

# Outline

1. Basic definitions and properties
2. Problem Statement
3. Fundamental Lemmas and Assumptions
4. Convergence Results for SG
5. Variance Reduction Techniques
6. Supplements



# Strongly convex $f$ and fixed step-size

## Theorem 1

For all  $k \in \mathbb{N}$  and constant step-size  $\alpha_k = \alpha$  satisfying

$$0 < \alpha \leq \frac{c}{LM_G}, \quad (5)$$

the expected optimality gap satisfies

$$\begin{aligned} \mathbb{E}[f(\mathbf{w}_k) - f^*] &\leq \frac{\alpha LM}{2\mu c} + (1 - \alpha\mu c)^{k-1} \left( f(\mathbf{w}_1) - f^* - \frac{\alpha LM}{2\mu c} \right) \\ &\xrightarrow{k \rightarrow \infty} \frac{\alpha LM}{2\mu c} \end{aligned} \quad (6)$$

where  $M_G = M_V + c_0^2$ .

If  $g(\mathbf{w}_k; \zeta_k)$  is unbiased estimate of  $\nabla f(\mathbf{w}_k)$ , then  $c = 1$ , we may assume  $M_G = 1$  and retrieve  $\alpha \in (0, 1/L]$  of GD

## Additional notes

$$\mathbb{E}[f(\mathbf{w}_k) - f^*] - \frac{\alpha LM}{2\mu c} \leq (1 - \alpha\mu c)^{k-1} \left( f(\mathbf{w}_1) - f^* - \frac{\alpha LM}{2\mu c} \right)$$

Fast convergence to a neighborhood of the optimal value, but noise in the gradient prevented further progress (convergence to an ambiguity ball)

Optimality gap  $\frac{\alpha LM}{2\mu c}$

Contraction constant after  $k$  iteration  $(1 - \alpha\mu c)^{k-1}$

A simple modification: run SG with a fixed step-size, and after convergence halve the step-size and run SG again, ...

- How  $E[f(\mathbf{w}_k)]$  against  $k$  behaves now?
- No sub-optimality gap
- Each time the step-size is cut in half, double the number of iterations are required
- **Effective convergence rate**  $\mathcal{O}(1/k)$ , why?

# Strongly convex $f$ and diminishing step-size

## Theorem 2

For all  $k \in \mathbb{N}$  and diminishing step-size  $\alpha_k$  satisfying

$$\alpha_k = \frac{\beta}{\gamma + k}, \text{ for some } \beta > \frac{1}{\mu c} \text{ and } \gamma > 0 \text{ s.t. } \alpha_1 \leq \frac{c}{LM_G},$$

the expected optimality gap satisfies

$$\mathbb{E}[f(\mathbf{w}_k) - f^*] \leq \frac{\nu}{\gamma + k} \quad (7)$$

where

$$\nu := \max \left\{ \frac{\beta^2 LM}{2(\beta \mu c - 1)}, (\gamma + 1)(f(\mathbf{w}_1) - f^*) \right\}$$

Usually first term of  $\nu$  determines the asymptotic convergence of  $\mathbb{E}[f(\mathbf{w}_k) - f^*]$

► Proof

## Additional notes

New step-size parameter  $\beta > \frac{1}{\mu c}$ :

Sensitive to overestimation of  $\mu$

higher  $\mu \rightarrow$  smaller  $\beta \rightarrow$  slower convergence rate

Constant step-size and mini-batch vs diminishing step-size

For mini-batch, define  $g(\mathbf{w}_k; \zeta_k) = \frac{1}{N_m} \sum_{i \in [N_m]} \nabla f_{\zeta_k, i}(\mathbf{w}_k)$

Mini-batch with small constant  $\alpha > 0$ ,

$$\mathbb{E}[f(\mathbf{w}_k) - f^*] \leq \frac{\alpha LM}{2\mu c N_m} + (1 - \alpha\mu c)^{k-1} \left( f(\mathbf{w}_1) - f^* - \frac{\alpha LM}{2\mu c N_m} \right)$$

Simple SG with small constant  $\alpha/N_m$ , (cheap iterations, many iterations)

$$\mathbb{E}[f(\mathbf{w}_k) - f^*] \leq \frac{\alpha LM}{2\mu c N_m} + \left( 1 - \frac{\alpha\mu c}{N_m} \right)^{k-1} \left( f(\mathbf{w}_1) - f^* - \frac{\alpha LM}{2\mu c N_m} \right)$$

# Convex $f$ and diminishing step-size

## ● Notations:

- $\mathbb{E}[g(\mathbf{w}; \zeta_k) | \mathbf{w}_k] \in \partial f(\mathbf{w}_k)$ : noisy unbiased sub-gradient of convex  $f$
- $f_{\text{best}}(\mathbf{w}_k) = \min(f(\mathbf{w}_1), \dots, f(\mathbf{w}_k))$
- $\mathbb{E}[\|g(\mathbf{w}_k; \zeta_k)\|_2^2] \leq G^2$  for all  $k$ , and  $\sup_{\mathbf{w} \in \mathcal{W}} \mathbb{E}[\|\mathbf{w}_1 - \mathbf{w}^*\|_2^2] \leq R^2$

## Theorem 3

Under some mild conditions and for square summable but not summable step-size, we have convergence in expectation

$$\mathbb{E}[f_{\text{best}}(\mathbf{w}_k) - f^*] \leq \frac{R^2 + G^2 \sum_{i \in [k]} \alpha_i^2}{2 \sum_{i \in [k]} \alpha_i}$$

and for any arbitrary  $\epsilon, \delta > 0$ , we have convergence in probability:

$$\Pr(f_{\text{best}}(\mathbf{w}_k) - f^* \geq \epsilon) \leq \delta$$

# Convex $f$ and diminishing step-size

## Theorem 4

For convex  $L$ -smooth function  $f$ , i.i.d. stochastic gradient of variance bound  $\sigma^2$ , and diminishing step-size  $\alpha_k = \frac{1}{L+\gamma^{-1}}$ , where  $\gamma = \frac{R}{G} \sqrt{\frac{2}{k}}$ , we have

$$\mathbb{E} \left[ f \left( \frac{1}{k} \sum_{i \in [k]} w_k \right) - f^* \right] \leq R \sqrt{\frac{2\sigma^2}{k}} + \frac{LR^2}{k} \quad (8)$$

Proof: see [Bubeck 2015, Theorem 6.3]

Improved gain for mini-batch of size  $N_m$ :  $\sigma^2 \rightarrow \sigma^2/N_m$

# Non-convex objective function

## Theorem 5

With fixed step-size as of (5), for all  $K \in \mathbb{N}$ , we have

$$\mathbb{E} \left[ \sum_{k \in [K]} \|\nabla f(\mathbf{w}_k)\|_2^2 \right] \leq \frac{K\alpha LM}{c} + \frac{2(f(\mathbf{w}_1) - f_{\inf})}{c\alpha} \quad (9)$$

and therefore

$$\mathbb{E} \left[ \frac{1}{K} \sum_{k \in [K]} \|\nabla f(\mathbf{w}_k)\|_2^2 \right] \leq \frac{\alpha LM}{c} + \frac{2(f(\mathbf{w}_1) - f_{\inf})}{Kc\alpha} \xrightarrow{K \rightarrow \infty} \frac{\alpha LM}{c} \quad (10)$$

*Proof:* Recursively  $\forall k \in [K]$ , take total expectation from (4), use (5), observe

$$f_{\inf} - f(\mathbf{w}_1) \leq \mathbb{E}[f(\mathbf{w}_{K+1})] - f(\mathbf{w}_1) \leq -\frac{1}{2}c\alpha \sum_{k \in [K]} \mathbb{E}[\|\nabla f(\mathbf{w}_k)\|_2^2] + \frac{1}{2}K\alpha^2 LM.$$

$f_{\inf}$  is not necessarily  $f^*$

SG spends increasingly more time in regions where the objective function has a “relatively” small gradient. Also usual tradeoff on step-size.

# Non-convex objective function

## Theorem 6

With square summable but not summable step-size, we have for any  $K \in \mathbb{N}$

$$\mathbb{E} \left[ \sum_{k \in [K]} \alpha_k \|\nabla f(\mathbf{w}_k)\|_2^2 \right] < \infty \quad (11)$$

and therefore

$$\mathbb{E} \left[ \frac{1}{\sum_{k \in [K]} \alpha_k} \sum_{k \in [K]} \alpha_k \|\nabla f(\mathbf{w}_k)\|_2^2 \right] \xrightarrow{K \rightarrow \infty} 0 \quad (12)$$

Proof: **HW 2.3!**

The expected gradient norm cannot stay bounded away from zero



# Foods for thought

1. Recall from Theorem 1 that SG with a constant step-size converges linearly to an ambiguity ball whose radius is determined by the variance of the gradient noise

Observe that taking  $N_k > 1$  samples “with replacement” implies multiplying the radius of the ambiguity ball by  $N_k^{-1}$  and conclude “doubling the batch size cuts the error in half”

Observe that taking  $N_k > 1$  samples “without replacement” implies multiplying the radius of the ambiguity ball by  $\frac{N - N_k}{N N_k}$

Modify the generic SG algorithm with a dynamic batch size.

Can we recover the linear convergence rate to  $w^*$ ? Linear in terms of iterations or workload (effect computations)? Note the increasing cost of iterations (due to larger  $N_k$  with  $k$ )

2. Often in practice, features (inputs,  $x \in \mathcal{X}$ ) of dimension  $d$  are very sparse (at most  $z \ll d$  non-zero elements)

Modify SG method to have  $\mathcal{O}(z)$  cost per iteration instead of the original  $\mathcal{O}(d)$

Can we do that for all objective functions? What about an SVM classifier?

3. In decentralized/distributed computing, we may have a high communication overhead to exchange  $w_k$  among workers. Can we use the vanilla SG method to tradeoff the costs between computation and communication?

# Outline

1. Basic definitions and properties
2. Problem Statement
3. Fundamental Lemmas and Assumptions
4. Convergence Results for SG
5. Variance Reduction Techniques
6. Supplements

# Stochastic variance reduced gradient (SVRG)

SVRG (Johnson&Zhang, 2013; Zhang et. al., 2013)

```
Inputs: Epoch length  $T$ , number of epochs  $K$   
for  $k = 1, 2, \dots, K$  do  
  Compute all gradients and store  $\tilde{\nabla} f := \frac{1}{N} \sum_{i \in [N]} \nabla f_i(\tilde{\mathbf{w}}_k)$   
  Initialize  $\mathbf{w}_{k,0} \leftarrow \tilde{\mathbf{w}}_k$   
  for  $t=1, \dots, T$  do  
    Sample  $\zeta_k$  uniformly from  $[N]$   
     $\mathbf{w}_{k,t} \leftarrow \mathbf{w}_{k,t-1} - \alpha_k \left( \nabla f_{\zeta_k}(\mathbf{w}_{k,t-1}) - \nabla f_{\zeta_k}(\tilde{\mathbf{w}}_k) + \tilde{\nabla} f \right)$   
  end for  
  Update  $\tilde{\mathbf{w}}_{k+1} \leftarrow \mathbf{w}_{k,T}$   
end for  
Return:  $\tilde{\mathbf{w}}_{K+1}$ 
```

- One memory, two gradients per inner loop
- **Linear convergence rate** (given a sufficiently large  $T$ )

► Proof

# Stochastic average gradient (SAG)

SAG (Schmidt&Le Roux&Bach, 2012, 2017)

**for**  $k = 1, 2, \dots$ , **do**

Sample  $\zeta_k$  uniformly from  $[N]$  and observe  $\nabla f_{\zeta_k}(\mathbf{w}_k)$

Update for all  $i \in [N]$ ,  $\hat{g}_i(\mathbf{w}_k) = \begin{cases} \nabla f_i(\mathbf{w}_k), & \text{if } i = \zeta_k \\ \hat{g}_i(\mathbf{w}_{k-1}), & \text{otherwise} \end{cases}$

Update  $\mathbf{w}_{k+1} \leftarrow \mathbf{w}_k - \frac{\alpha_k}{N} \sum_{i \in [N]} \hat{g}_i(\mathbf{w}_k)$

**end for**

- Almost same convergence rate (and same proof) as of SVRG
- A memory of size  $N$
- Biased gradient estimates:  $\mathbb{E} \left[ \frac{1}{N} \sum_{i \in [N]} \hat{g}_i(\mathbf{w}_k) \right] = \frac{1}{N} \sum_{i \in [N]} \nabla f_i(\mathbf{w}_k)$  does not hold necessarily
- Table averaging representation and SAG<sup>A</sup> extension

# Which algorithm to choose?

## CA1: Closed-form solution vs iterative approaches

Consider  $\mathbf{x}^* = \underset{\mathbf{w} \in \mathbb{R}^d}{\text{minimize}} \frac{1}{N} \sum_{i \in [N]} \|\mathbf{w}^T \mathbf{x}_i - \mathbf{y}_i\|^2 + \lambda \|\mathbf{w}\|_2^2$  for dataset  $\{(\mathbf{x}_i, \mathbf{y}_i)\}$

- 1) Find a closed-form solution for this problem
- 2) Consider “Individual household electric power consumption” dataset ( $N = 2075259$ ,  $d = 9$ ) and find the optimal linear regressor from the closed-form expression
- 3) Repeat 2) for “Greenhouse gas observing network” dataset ( $N = 2921$ ,  $d = 5232$ ) and observe the scalability issue of the closed-form expression
- 4) How would you address even bigger datasets?

## CA2: Deterministic/stochastic algorithms in practice

Consider logistic ridge regression  $f(\mathbf{w}) = \frac{1}{N} \sum_{i \in [N]} f_i(\mathbf{w}) + \lambda \|\mathbf{w}\|_2^2$  where  $f_i(\mathbf{w}) = \log(1 + \exp\{-\mathbf{y}_i \mathbf{w}^T \mathbf{x}_i\})$  for “Greenhouse gas observing network” dataset

- 1) Solve the optimization problem using GD, stochastic GD, SVRG, and SAG
- 2) Tune a bit hyper-parameters (including  $\lambda$ )
- 3) Compare these solvers in terms complexity of hyper-parameter tuning, convergence time, convergence rate (in terms of # outer-loop iterations), and memory requirement

## Some references

- L. Bottou, F. E. Curtis, and J. Nocedal, "Optimization methods for large-scale machine learning," SIAM Review, 2018.
- S. Bubeck, "Convex optimization: Algorithms and complexity," Foundations and Trends in Machine Learning, 2015.
- S. Boyd and A. Mutapcic, "Stochastic subgradient methods," Lecture Notes for EE364b, Stanford University, 2018.
- R. Johnson and T. Zhang, "Accelerating stochastic gradient descent using predictive variance reduction," NIPS, 2013.
- L. Zhang, M. Mahdavi, and R. Jin, "Linear convergence with condition number independent access of full gradients," NIPS 2013.
- M. Schmidt, N. Le Roux, and F. Bach, "Minimizing finite sums with the stochastic average gradient," Mathematical Programming, 2017.

# Outline

1. Basic definitions and properties
2. Problem Statement
3. Fundamental Lemmas and Assumptions
4. Convergence Results for SG
5. Variance Reduction Techniques
6. Supplements

# Proof sketch for Theorem 1

Use (4), Polyak-Lojasiewicz inequality (a consequence of strong convexity) and (5), and observe that

$$\begin{aligned}\mathbb{E}_{\zeta_k}[f(\mathbf{w}_{k+1})] - f(\mathbf{w}_k) &\leq -\left(c - \frac{1}{2}\alpha LM_G\right) \alpha \|\nabla f(\mathbf{w}_k)\|_2^2 + \frac{1}{2}\alpha^2 LM \\ &\leq -\frac{1}{2}\alpha c \|\nabla f(\mathbf{w}_k)\|_2^2 + \frac{1}{2}\alpha^2 LM \\ &\leq -\alpha\mu c (f(\mathbf{w}_k) - f^\star) + \frac{1}{2}\alpha^2 LM\end{aligned}$$

Subtract  $f^\star$  from both sides, take total expectation, and rearrange:

$$\mathbb{E}[f(\mathbf{w}_{k+1}) - f^\star] \leq (1 - \alpha\mu c) \mathbb{E}[f(\mathbf{w}_k) - f^\star] + \frac{1}{2}\alpha^2 LM$$

Make it a contraction inequality (as  $0 < \alpha\mu c \leq \frac{\mu c^2}{LM_G} \leq \frac{\mu}{L} \leq 1$ )

$$\mathbb{E}[f(\mathbf{w}_{k+1}) - f^\star] - \frac{\alpha LM}{2\mu c} \leq (1 - \alpha\mu c) \left( \mathbb{E}[f(\mathbf{w}_k) - f^\star] - \frac{\alpha LM}{2\mu c} \right).$$



## Proof sketch for Theorem 2

First observe that  $\alpha_k LM_G \leq \alpha_1 LM_G \leq c$ . Use (4) and Polyak-Lojasiewicz inequality and show that

$$\mathbb{E}_{\zeta_k} [f(\mathbf{w}_{k+1})] - f(\mathbf{w}_k) \leq -\alpha_k \mu c (f(\mathbf{w}_k) - f^*) + \frac{1}{2} \alpha_k^2 LM$$

Subtract  $f^*$  from both sides, take total expectation, and rearrange:

$$\mathbb{E} [f(\mathbf{w}_{k+1}) - f^*] \leq (1 - \alpha_k \mu c) \mathbb{E} [f(\mathbf{w}_k) - f^*] + \frac{1}{2} \alpha_k^2 LM$$

Now prove by induction and use inequality  $k^2 \geq (k+1)(k-1)$

► Return

## Proof sketch for Theorem 3

Use convexity of  $f$  ( $f^\star - f(\mathbf{w}_k) \geq \mathbb{E}[g(\mathbf{w}; \zeta_k) | \mathbf{w}_k]^T (\mathbf{w}^\star - \mathbf{w}_k)$ ) to show

$$\mathbb{E} [\|\mathbf{w}_{k+1} - \mathbf{w}^\star\|_2^2 | \mathbf{w}_k] \leq \|\mathbf{w}_k - \mathbf{w}^\star\|_2^2 - 2\alpha_k (f(\mathbf{w}_k) - f^\star) + \alpha_k^2 G^2$$

Take expectation and apply recursively to show

$$\mathbb{E} [\|\mathbf{w}_{k+1} - \mathbf{w}^\star\|_2^2] \leq \mathbb{E} [\|\mathbf{w}_1 - \mathbf{w}^\star\|_2^2] - 2 \sum_{i \in [k]} \alpha_i (\mathbb{E}[f(\mathbf{w}_i)] - f^\star) + G^2 \sum_{i \in [k]} \alpha_i^2$$

Conclude that for square summable but not summable step-size,  $\min_{i \in [k]} \mathbb{E}[f(\mathbf{w}_i)] \rightarrow f^\star$

Use Jensen's inequality and concavity of minimum to show convergence in expectation  $\mathbb{E}[f_{\text{best}}(\mathbf{w}_k)] = \mathbb{E}[\min_{i \in [k]} f(\mathbf{w}_i)] \leq \min_{i \in [k]} \mathbb{E}[f(\mathbf{w}_i)] \rightarrow f^\star$

Use Markov's inequality to show convergence in probability:

$$\Pr(f_{\text{best}}(\mathbf{w}_k) - f^\star \geq \epsilon) \leq \frac{\mathbb{E}[f_{\text{best}}(\mathbf{w}_k) - f^\star]}{\epsilon}$$

# Linear convergence of SVRG

Variance decomposition:

$$\mathbb{E} [\|\mathbf{w} - \mathbb{E} [\mathbf{w}] \|_2^2] \leq \mathbb{E} [\|\mathbf{w}\|_2^2] - \|\mathbb{E} [\mathbf{w}] \|_2^2 \leq \mathbb{E} [\|\mathbf{w}\|_2^2]$$

Show

$$\mathbb{E}_{\zeta_k} \left[ \left\| \nabla f_{\zeta_k}(\mathbf{w}_{k,t-1}) - \nabla f_{\zeta_k}(\tilde{\mathbf{w}}_k) + \tilde{\nabla} f \right\|_2^2 \right] \leq 4L (f(\mathbf{w}_{k,t-1}) + f(\tilde{\mathbf{w}}_k) - 2f^*)$$

Use the inner-loop iteration and bound  $\mathbb{E}_{\zeta_k} [\|\mathbf{w}_{k,t} - \mathbf{w}^*\|_2^2]$ . You may need to use convexity of  $f$

Sum  $\mathbb{E}_{\zeta_k} [\|\mathbf{w}_{k,t} - \mathbf{w}^*\|_2^2]$  over the inner loop ( $t \in [T]$ ) and cancel some terms from both sides

Show and use for every outer iteration to observe the linear convergence rate: if  $a < ba + c$  for  $b \in (0, 1)$ , then

$$a - \frac{c}{1-b} \leq b \left( a - \frac{c}{1-b} \right)$$